PROGRAMMING THE POPULATION CENSUS*

By: Richard A. Hornseth, Bureau of the Census

A brief description of the over-all plans for processing returns from the 1960 Census of Population and Housing will provide a perspective for a discussion of the uses of the computer in processing the Census and the associated tasks of programming. Processing begins with the delivery to our Jeffersonville Office of enumeration books which consist of FOSDIC documents permitting the optical-electronic translation of enumerator entries to pulses on magnetic tape. At Jeffersonville, the books are checked in by enumeration district, or ED, assembled by counties and large cities within States, provided with ED numbers and population and housing field counts in FOSDIC form, and then microfilmed. Books for the 25 percent sample enumeration are handled separately from those for the 100 percent enumeration and require, before microfilming, an additional step of manual coding in FOSDIC form of certain written entries such as occupation. After the microfilm is developed, it is sent to Washington where the translation to magnetic tape is performed by FOSDIC.

The FOSDIC output tapes serve as initial input to the computer where each record for persons and housing units is edited for blank items and inconsistencies and a corrected version of it, for the sample, is put out on tape. The 100 percent record, however, is tallied according to the categories to be shown in publication immediately after the editing for blank and inconsistent entries, and only the accumulated tallies for each ED are put on tape. In the editing process, a diary summarizing and evaluating the quality of the records and comparing field and computer counts for each ED is prepared to determine which ED's require correction outside the computer and what types of correction are called for. Also, in the editing process, a control tape listing all ED numbers for a State for purposes of control is checked off to account for each ED. In the 100 percent operation, final population and housing unit counts for each ED are inserted on the control tape and a quick computer pass of the completed control tape will provide the State population totals required for certification to the President of the United States. In addition, the control tape, when merged with a master ED identification tape providing publication area codes for each ED, permits the preparation of the Population Series P-A publication which shows the number of inhabitants for every area in the United States. In the sample operation, the control tape is posted with ratio estimate counts required for later determination of weights for the sample records.

The output of the first computer pass in the 100 percent operation consists of a set of about 500 counts or tallies for each ED of persons and housing units by the characteristics appearing in all tables for the 100 percent publications: Series P-B for population, block statistics for housing, tract statistics for population and housing, and other reports. A merge of this ED tally output with the master ED identification tape permits the summarization of tallies to every area or type of area shown in the publications. These summaries, when merged with tapes containing place names and historical detail, enable the assembly of tables into publication format for running on the High Speed Printer to provide copy for photo-offset reproduction.

The output of the first computer pass in the sample operation is a partially edited file of records for persons and housing units. Another pass of this file is required to complete the editing and to assign the sample weights to each record. The weight for a person or unit in a ratio estimate category, for example, male renter heads 25 to 44 years, is roughly the ratio of the 100 percent count in that category to the corresponding sample count over a collection of ED's comprising the smallest area of publication. The ratios are determined by the computer in an operation involving the merge of the sample control tape containing the sample ratio estimate counts with a tape containing corresponding 100 percent counts. The completely edited and weighted sample file of individual person and housing records is then passed through the computer as many times as required, using sub-files when necessary, to produce all sample tabulations. Finally, these tabulations are summarized and assembled into publication tables for High Speed Printer photo-offset copy.

The above sketch has indicated or implied the major uses planned for the computer: Editing of records; check-off or control; evaluation of quality; determination of sample weights; tabulation; preparation of publication copy. Several subsidiary functions not mentioned include the preparation of the master ED identification tape, historical data and place name tapes, listings of ED numbers for field and processing control outside the computer, and progress reports.

Mainly, the many uses planned for the computer arise from experienced-based expectations of, and sometimes merely hopes for, gains in quality, time, control, economy, or convenience, and these uses have been arrived at somewhat piecemeal as each processing problem came under consideration. Our British counterparts have not been so exposed to large-scale digital computers as we, and they presently plan a much more modest use for computers in their coming census. In one manner or another, we have been led to use the computer as much as possible and the upshot is that we quite likely may become able theoretically to process completely the 100 percent returns for an average State from original books to publication copy in something under a calendar month with a directly involved staff measuring perhaps less than one hundred man-months. In actual production, a longer calendar time may be required for a particular State since its separate operations must be scheduled among those required for other States. We do expect to complete the bulk of the 100 percent publication program for the 50 States by mid-1961.

This particular achievement in automation stands to be accomplished without it or its implications having been foremost in mind, at least among those concerned with the detailed planning. In fact, at the programming level, the all-absorbing problem is how to accomplish a particular task within the limits of the equipment and experience with it, marvelous though that equipment may be. The day-by-day accretion of experience and skill in the use of equipment at hand provides the basis for accomplishment, not the presence of equipment or the appeal of an automation goal. Thus, programmers are apt to be cautious and tend to resist grandiose push button plans. Each particular accomplishment too painfully has been earned.

A closer look at several of the major functions planned for the computers will illustrate the problems in planning specific computer operations and the requirements of balance among such considerations as limitations of the computer and auxiliary equipment, availability of computer time, supply of experienced programming staff, pressure of deadlines, desires for economy and appeals for extra quality, detail, or information--all weighed on the basis of experience gained since 1950.

Tabulation, of course, is the prime function of the computers and it would appear that tabulation particularly would be facilitated by them. Our present computers, for example, can produce a 5,000 cell tabulation at the rate of about 3,000 persons per minute. The potential tabulations of census material are enormous and, unfortunately, it seems that every possible tabulation can claim a user somewhere. Yet, the content and coverage of publications planned for 1960 represent perhaps only about a 25 percent increase over that for 1950. The major limiting factors are economic, mainly the costs of printing and the availability of computer time within the census period. The Population Series P-D tabulations alone may require something like six months of computer time at the rate of 1,000 available hours per month. This can be construed really as a consequence of a limitation of computer capacity. Series P-D would require about 100,000 counters if it were to be tabulated in one pass of the sample file. However, our computers can accommodate only 5,000 counters easily and up to 15,000 with some difficulty. Hence, something of the order of ten passes of the sample file may be necessary.

Printing costs present a particularly stringent limit to publication plans and thus have initiated considerable exploration and expansion of computer use into what may be called the typesetting field. Preparation of tables for photo-offset reproduction from High Speed Printer copy is planned for practically all publication for both the 100 percent and sample results. Not only reductions in cost but gains in time are expected. Table preparation does not involve significant amounts of computer time, but it imposes extraordinary demands on the ingenuity, skill, and regard for detail of programmers. The relatively limited capacity of the computer and the severe limitations of the High Speed Printer pose problems in programming of such magnitude that computer table preparation is not being attempted for the Population Series P-A publication and the United States Summaries. Series P-A involves quantities of historical and footnote material that can best be prepared by hand. The United States Summaries require table widths not possible on the High Speed Printer.

Our experience has been limited in the field of table preparation by the computer. We are proceeding to some extent on the feeling that the programming time and effort expended on preparing tabulations in the computer in a form permitting ease of posting for a manual preparation of printer copy could be extended without too great an effort to programming full table preparation in the computer. In any event, considerations of publication costs and timing are overriding and impel us to use the computer as much as possible.

A discussion of one more major function of computer processing will suffice. This concerns the editing of individual records for completeness of entry and consistency. The main considerations again are costs and time. A clerical editing operation is costly and time consuming and, in some quarters, has not been regarded as particularly effective. In fact, many feel that a computer supplied with sufficient rules can do a better job of making assignments for, say, unknown age or unknown income. The problem revolves around what constitutes sufficient rules. Rules that anticipate all conceivable possibilities generally exceed the capacity of the computer or tie up more programming talent than can be afforded for computer coding that may apply to only a few cases in the entire census. Consequently, editing rules used in the computer are being pruned to those considered essential for maintaining the quality of the census. For protection from the unexpected and the rare violations of quality, a diary for each ED, produced during the editing pass, will provide for inspection of counts of the number of imputations by type, and, when the counts exceed certain levels, the ED's will be flagged for manual inspection of the original books or their microfilm. In this fashion, it is expected that enumeration books can be processed with a minimum of manual intervention and without a clerical edit other than that provided in the field quality control sample check. The quality of the edit will not suffer. Computer capacity does provide for fairly complicated editing rules, and for handling of an entire household as a unit in both the 100 percent and sample operations. Also, new editing techniques are made possible. For example, income for a person not reporting it can be made the same as that of the last person previously processed having the same characteristics of age, sex, weeks worked, and occupation. Though the main considerations in applying the computer to the problem of editing records have been cost and time, improvement over 1950 in quality of the edit can be assured.

The use of computers in processing returns from the 1960 Census represents an extensive development in the application of experience with large-scale digital computers which the Bureau has had for the past 10 years. The presently planned use of computers ought to produce results equal to or better than those produced from the 1950 Census by more traditional means and certainly much more quickly and economically. There yet remain limitations of equipment, personnel, and experience, which will prevent a full realization of a goal of automatic data processing, if any of us ever had that goal. In the decade after the census, the materials and experience gained during the census should support such a surge of development in automatic data processing that the 1970 Census processing operation may become truly a push button affair. Those who do not consider this a proper goal can take heart in the reasonable expectation that marked improvements in quality, utility, and timeliness generally accompany the attainment of such a goal. They may take comfort in the knowledge that at the detailed planning stages the concern is with computer application to specific problems involving quality and utility. At that level, the computer is not considered a means for short cut but a remarkable device for solving problems.

*Paper presented at the session December 29, 1959, on "How the 1960 Census will be taken" of the Social Statistics Section of the American Statistical Association at its Annual Meeting in Washington, D. C.